REVIEW 10.1111/j.1469-0691.2009.02962.x

How will next-generation sequencing contribute to the knowledge concerning *Helicobacter pylori*?

L. Engstrand

Department of Bacteriology, Swedish Institute for Infectious Disease Control & Department of Microbiology, Tumor and Cell Biology, Karolinska Institutet, Stockholm, Sweden

Abstract

Molecular microbiology has revolutionized the landscape of microbiology and will continue to do so by providing new solutions for microbe identification and characterization. This applies also to the study of *Helicobacter pylori* where current genotypic (molecular) methods are important complements or alternatives to phenotypic methods. Besides providing sensitivity and specificity and an enhancement of the detection process, they also reduce much of the subjectivity inherent in the interpretation of morphological and biological data. Another key advantage of molecular methods is that they allow the identification of novel virulence factors of pathogenic bacteria. For example, such gene products enable *H. pylori* to establish itself within the gastric environment and enhance its potential to cause disease. Next-generation sequencing will open up new areas of research for those involved in the field of *Helicobacter* research and will also provide information that will help to develop novel treatment strategies and increase our understanding of the mechanisms behind chronic inflammations in the gut. The analysis of data resulting from a large-scale sequencing project requires the use of bioinformatics, including standard BLAST analysis, annotation or clustering, and assembly competence. However, the amount of data produced by the next-generation sequencing platforms will require a bioinformatics capacity at the industrial scale, which may limit the availability of such technologies. Consequently, building effective new approaches to data analysis must be given high priority.

Keywords: Bioinformatics, data analysis, *Helicobacter pylori*, next-generation sequencing, technology platforms
Clin Microbiol Infect 2009; **15:** 823–828

Corresponding author and reprint requests: L. Engstrand, Department of Bacteriology, Swedish Institute for Infectious Disease Control & Department of Microbiology, Tumor and Cell Biology, Karolinska Institutet, Stockholm, Sweden
E-mail: lars.engstrand@smi.se

Technology Platforms

Growing demand in both the research and clinical markets has fueled the development of more efficient genomic sequencing methods [1,2]. Such methods are already several orders of magnitude more efficient than the Sanger capillary-array electrophoresis machines that were used in the human genome project. Massively parallel DNA sequencing platforms have not only reduced the cost of DNA sequencing, but also have moved the technology from major genome centres to individual investigators. The new platforms will dramatically accelerate biological and biomedical research, by enabling the comprehensive analysis of genomes to become inexpensive, routine and widespread. Below, three commercial systems are briefly described (Table 1) and each of them has the potential to contribute to our knowledge concerning *Helicobacter pylori*.

Roche/454 FLX pyrosequencer

Multiple whole prokaryote genomes can easily be sequenced using the 454 FLX system (Roche Diagnostics, Basel, Switzerland.) [3,4]. This high-throughput technology that sequences in real time provides long reads (400 bp) that facilitate the completion of near-finished draft sequences in a single instrument run. Large-size genomic DNA samples are randomly fragmented into small 300- to 800-bp fragments for shotgun sequencing. Addition of adapters to the fragments creates a library of DNA fragments, which is immobilized on DNA capture beads, whereafter PCR amplification takes place in water-in-oil microreactors, resulting in millions of copies of the template. Finally, the microreactor is broken and beads

TABLE 1. Comparison of next-generation sequencing technologies

	454 FLX	Solexa	SOLiD
Read length	250–400 bp	25–35 bp	25–35 bp
Reads	1 M	30 M	90 M
Data	400 Mb	3 Gb	30 Gb
Scale-up of number of reads	+	+++	+++
Future increase of read length	++	+	+
Access to instruments	+++	+++	+++
Drawbacks	High error rate for homopolymers	Error rate increases with read length	Error rate increases with read length
Advantages	Long read length	Easy to scale up	Easy to scale up

carrying single-stranded DNA templates are individually sequenced on a picotitre plate device. The generated sequences are then assembled into a number of unordered contigs using specific assembler software. Finally, a consensus sequence is generated.

The sequencing depth achieved with 454 FLX titanium sequencing systems ensures the accurate characterization of microbial or bacterial diversity, the sensitive detection of even rare mutations, and the rapid discovery of the disease-causing agents [5]. Furthermore, this system for ultra-high-throughput DNA sequencing is used for *de novo* sequencing and resequencing of genomes, for metagenomics, and for targeted sequencing of DNA regions of interest. The newest version generates up to 400 million bases per 10-h instrument run. The key advantage of this technology is read-length (up to 400 bp, which is necessary in *de novo* assembly and metagenomics). However, a major limitation is that no prevention of multiple incorporations at a given cycle is provided, which leads to homopolymer errors.

The technology has enabled a number of peer-reviewed studies in diverse research fields, such as cancer and infectious diseases, drug discovery, marine biology, anthropology, paleontology, and many more. The value of the FLX System for bacterial sequencing applications is emphasized by a number of important studies, including a study of *Mycobacterium tuberculosis* that resulted in the identification of the first tuberculosis-specific drug candidate in 40 years [6]. FLX System pyrosequencing has so far been the method of choice for sequencing of *H. pylori* and for exploring the human stomach microbiota [7,8].

Illumina/Solexa genome analyzer

Illumina sequencing technology (Illumina Inc., San Diego, CA, USA), or the Solexa platform, allows for the selection of any single-nucleotide polymorphism or probe, enabling dense, uniform coverage across the genome and the ability to target any genomic region [9,10]. This platform is based on massively parallel sequencing of millions of fragments using a reversible terminator-based sequencing chemistry. The technology, together with a software application, allows a scalable system that many consider cost-effective and accurate. It relies on the attachment of randomly fragmented genomic DNA to an optically transparent surface. Attached DNA fragments are extended and subjected to bridge amplification to create an ultra-high-density sequencing flow cell with ≥50 million clusters, each containing approximately 1000 copies of the same template. These templates are sequenced using a four-colour DNA sequencing-by-synthesis technology that employs reversible terminators with removable fluorescent dyes. This approach ensures high accuracy and true base-by-base sequencing, eliminating sequence context-specific errors and enabling sequencing through repetitive sequences. After completion of the first read, the templates can be regenerated *in situ* to enable a second >36-bp read from the opposite end of the fragments. A paired-end module directs the regeneration and amplification operations to prepare the templates for the second round of sequencing. Once the original templates are cleaved and removed, the reverse strands undergo sequencing-by-synthesis. The second round of sequencing occurs at the opposite end of the templates, generating >36 bp reads for a total of >3 Gb of data, which is an obvious advantage when sequencing large genomes. The short read-length (35 bp) is a limitation but, compared with 454 FLX pyrosequencing, homopolymer errors are less of an issue with this technology.

Applied Biosystems SOLiD™ system

Sequencing by ligation generates DNA by measuring the serial ligation of an oligonucleotide. This technology is used in the SOLiD system (Applied Biosystems, Foster City, CA, USA) [11,12]. All fluorescently labelled oligonucleotide probes are present simultaneously and compete for incorporation. After each ligation, the fluorescence signal is measured and then cleaved before another round of ligation takes place. The SOLiD system is a massively parallel genomic analysis platform that supports a wide range of applications. The flexibility of two independent flow cells allows multiple experiments in a single run. The SOLiD system can cost effectively complete large-scale sequencing and, with a reference sequence for a microorganism, it is possible to

perform comparative sequencing or re-sequencing to characterize the genetic diversity within the organism's species or between closely-related species. The throughput is greater than 30 GB per run, although the read-length is limited to 35 bp.

Genome-wide Assessment of *H. pylori*: Insight into the *Helicobacter* Supragenome

The next-generation DNA sequencing platforms will soon be available in many research laboratories, which will allow researchers interested in *H. pylori* evolution, pathogenicity, colonization and other aspects to perform sequencing projects at the whole genome level. Comparative genomic analyses of many *H. pylori* strains will, for example, provide insights into the *Helicobacter* supragenome. Currently, we have some knowledge about the genomic plasticity of the *H. pylori* species; this genetic diversity calls attention to the need for markers of human virulence phenotypes and highlights the potential difficulty associated with this task. *H. pylori* strains are presently categorized according to certain proposed virulence-associated genes such as *cagA* and other genes in the *cag*PAI group. Additional potential virulence factors of *H. pylori* are so far unknown [13,14].

Only 10% of *H. pylori* infections are associated with severe pathology, including chronic gastritis, peptic ulcer, gastric adenocarcinoma and gastric mucosa-associated lymphoid tissue (MALT) lymphoma. Evolution of the bacterial infection towards malignancy only occurs in approximately 1% of the infected population, emphasizing the importance of markers that are able to discriminate between benign and pathogenic infections at an early stage. Numerous studies have focused on the identification of *H. pylori* genes associated with pathogenicity and such genes have served as virulence biomarkers. The current biomarkers fail, however, to discriminate among various *H. pylori*-related diseases and the importance of identifying such biomarkers for a given patient is controversial [15].

Comparative analyses of the first two Western *H. pylori* genome sequences (J99 isolated from a peptic ulcer patient [16] and 26695 [17] isolated from a patient with gastritis) revealed a significant macrodiversity (i.e. the presence or absence of the genes) and microdiversity (polymorphisms) [18]. The macrodiversity of *H. pylori* isolates was further illustrated by the comparative genome analysis of 15 Caucasian isolates, in a study that aimed to identify the core *H. pylori* genome and the strain-specific genes [19]. This comparison allowed the fraction of strain-specific genes to be extended from 6–7% (as determined by the comparison of the first two sequenced genomes) to 18–22%. More recently, Gressmann *et al.* [20] conducted a large study of 56 *H. pylori* and four *Helicobacter acinonychis* strains using whole genome microarrays. By extrapolation, it was estimated that the core *H. pylori* genome contains 1111 genes and that a weighted average of 27% of the genome is variably present in different isolates [20]. The whole genome array approach is limited to published genes present on the array and does not provide information about unknown genes in clinical isolates.

The next-generation sequencing platforms will avoid this limitation. We will have the opportunity to compare the genomes of a large number of isolates from specific gastroduodenal diseases aiming to obtain knowledge of *H. pylori* genes uniquely associated with these diseases (i.e. more strain-specific genes will be discovered). The data will extend our knowledge of the *H. pylori* core genome and supragenome (the collective genome of all *H. pylori* strains) and thus simplify the identification of strain-specific markers.

Another potential field of application is to extend the multilocus sequence typing approach to whole genome sequencing of *H. pylori* in studies of human migration. Now that *H. pylori* has become a marker of human prehistoric migration, whole genome sequencing of representative strains of the bacteria will probably increase our understanding of human expansion and help to determine when different parts of the world were peopled [21].

A large amount of intraspecies genetic variation has been observed for several bacteria. Analysis of 13 *Haemophilus influenzae* genomes showed that only approximately 50% of the genes were conserved among all strains [22]. This study supports the distributed-genome hypothesis, which states that the full complement of genes available to a given species exists in a supragenome pool, comprising one that each member of a population of naturally transformable bacterial strains contributes to and draws from, resulting in a high degree of genic diversity [23]. In another study, the genomes of eight clinical *Streptococcus pneumoniae* isolates were sequenced by 454 FLX pyrosequencing and combined with sequences from nine additional publicly available *S. pneumoniae* strains [24]. A global comparative analysis of the genes and genomes was performed and demonstrated a great genetic diversity among the strains. This total of 17 genomes contained orthologues shared among all strains (core genes), orthologues shared only between subsets of two or more strains (distributed genes), and genes unique to one strain. Genes from all strains were grouped into orthologous clusters and divided into the three categories.

The study by Hiller *et al.* [24] revealed that *S. pneumoniae* has a supragenome much larger than the genome of any individual strain. A mathematical model was also used to calculate the number of strains that needs to be sequenced to determine

the *S. pneumoniae* supragenome. The calculated number of strains required to identify 90% of the pneumococcal supragenome was 142. Hiller *et al.* [24] concluded that the pneumococcal diversity must lead to caution in the use of model strains to test and develop vaccines and drugs because effective targets in one strain may be missing in a significant percentage of the other strains. They speculated about the possibility that these bacteria have evolved multiple and redundant mechanisms to evade immunity and adapt to variations among hosts and their commensal microbiota. A similar evolvement can be attributed to *H. pylori* and, by using the next-generation sequencing platforms, we will probably be aware of this situation in the context of vaccine and drug development for *H. pylori* infection.

A similar approach to gaining insight into the *H. pylori* genome is a natural step for people in this field of research. One possibility is to use the 454 FLX pyrosequencing platform that allows for long reads (up to 400 bp) and a sequencing depth that ensures accurate whole genome sequencing of up to 25 *H. pylori* isolates per instrument run. Various *H. pylori*-associated diseases such as gastric MALT lymphoma, atrophic gastritis, gastric adenocarcinoma and duodenal ulcers should be included in these studies, together with control strains from asymptomatic individuals with chronic gastritis. By using sample-specific adaptors ligated onto each DNA fragment in a library sample, the GS FLX analysis software will recognize each adaptor, allowing for automated sorting and analysis of sample-specific reads.

We have used a similar approach to investigate the microbial contents in the human throat, stomach and in faecal samples [8]. DNA was extracted from each sample and the 16S rRNA gene contents were amplified by PCR using universal 16S primers. By using short sample-specific sequence tags, incorporated during the initial 16S PCR reactions, each sequence obtained on the picotitre plate was traced back to its original sample (i.e. a variant of the adaptor ligation approach above). We found an outstanding adaptation of *H. pylori* to the gastric environment in stomach samples that were positive for *H. pylori.* This bacterium dominated totally (accounting for >93% of the reads) in the diverse microbial communities of the investigated stomachs. The 454 FLX pyrosequencing platform allowed us to explore the gut microbiota with a sequencing depth that ensures an accurate characterization. Similar approaches for analyses of the gut microbiota have been reported recently [25–27].

By using gaskets, which physically divide the picotitre plate device into two smaller regions, 12 *H. pylori* DNA libraries can be sequenced on each side (450–650 000 reads per region). Based on the size of the *H. pylori* genome (1.6 Mb), one can estimate that this approach will generate an eight- to ten-fold greater coverage. By using mathematical models similar to those used for *S. pneumoniae*, it should be possible to estimate the number of whole genome-sequenced *H. pylori* strains required to define the supragenome of the bacteria [24]. The sequencing data from 25 *H. pylori* strains and the four previously published genomes (J99, 26695, HPAG1 and G27) [7,16,17,28] can then be compared to identify common, variable and strain-specific genes. To reach statistical significance, a power calculation is required, and such studies should only be performed on epidemiologically well-designed material representing true cases and controls [29]. Thus, strain specific genes can then be investigated to identify potentially novel virulence genes. The meta-genome data from such studies will also serve as a template for the design of microarrays, which will facilitate comparative genome hybridization experiments of a significantly larger number of strains. The expected outcome of such studies includes the identification of novel biomarkers that could potentially be used to classify *H. pylori* infections in a clinical setting.

Identifying Compensatory Mutations in Resistant *H. pylori*

Another study approach for next-generation sequencing comprises comparative genomics in *H. pylori* strains before and after treatment failures (i.e. when *H. pylori* develops resistance to the drug). We know that the correlation between the volume of drugs used and resistance development, together with the observation that resistance determinants often confer a fitness cost, has led to the question: is the development of antibiotic resistance reversible?

In an environment containing antibiotics, possession of the corresponding resistance gene confers a benefit to the organism. However, in a drug-free environment, resistant genotypes might have a lower fitness (i.e. reproductive ability) than their sensitive counterparts [30]. The reproductive ability is affected by several parameters, such as the relative rates at which susceptible and resistant bacteria grow in their hosts and in the environment, and how this affects clearance and transmission rates.

The fitness cost that chromosomal resistance mutations and plasmids often confer to the cells makes them less competitive than susceptible strains in an antibiotic-free environment. Counteracting this is the selection for a decreased cost of resistance. This can be accomplished by reversion of the resistance mutation or acquisition of a second-site compensatory mutation, restoring fitness close to or equal to the susceptible strains. One can conclude from several *in vitro* investigations that, under most conditions, compensation

rather then reversion is a more frequent event. This is because the mutational target for compensation is larger than for reversion.

From the perspective of reversibility of antibiotic resistance, compensation represents a substantial problem. This is because compensatory mutations most often do not reduce resistance levels or affect the resistance mechanism when restoring fitness. This influences the time required to reduce the population of resistant bacteria, which is inversely proportional to the cost of resistance. If the fitness cost of a resistance mutation is reduced to 1% by compensatory events, it will take 20 times longer to eliminate the compensated population (by the take-over from a susceptible clone) than it would have taken for the uncompensated resistant population with 20% lower fitness.

To assess the probability of reversibility of antibiotic resistance, it is vital to determine the fitness effect of resistance and compensatory mechanism in clinical isolates [31]. The problem with conducting studies in clinical isolates is the lack of functioning genetic tools (i.e. phages, transposomes and plasmids), if they exist at all for the organism in question, making these types of investigations basically difficult to conduct. However, one way of overcoming the problem is to apply a whole genome sequencing approach, with high-throughput sequencing techniques (454 FLX pyrosequencing, Solexa or SOLiD), to identify potential resistance and compensatory mechanisms.

The extensive strain material existing for *H. pylori*, from both before and after (failed) antibiotic treatment in the same patient, in combination with its chronic and clonal colonization profile, makes *H. pylori* an ideal candidate for assessing the fitness cost of resistance and for the identification of potential compensatory mechanisms. These types of studies have to forego any broadscale reversibility intervention for a specific antibiotic, which is also a fundamental requirement for assessing the risk of a worldwide eradication policy for *H. pylori*.

We have successfully sequenced four *H. pylori* isolates obtained from two patients before and after treatment using the 454 FLX pyrosequencer. The resistant isolates obtained after treatment showed different fitness in a mouse model [31] and, by employing the 454 FLX approach, we have identified a number of potential targets for fitness compensation and narrowed these down to a few targets (manuscript in preparation).

Considerations and Challenges

The next-generation DNA sequencing platforms will, without doubt, be applied to a variety of goals within the *H. pylori* research field. A number of reference genomes have been available for *H. pylori* for several years and the next possibility will be to resequence, perhaps not all the *H. pylori* genomes, but the regions of the genome to which a disease phenotype has been mapped. One example is to understand the adaptations of *H. pylori* to a persistent infection. We generated and finished the whole genome sequence for strain HPAG1 [7] using traditional Sanger sequencing with a ABI 3730xl capillary sequencer and, independently, using 454 FLX pyrosequencing. With both technologies, we identified 43 genes that were not detectable in the previously sequenced Hp 26695 or J99 strains from patients with acid-peptic disease, and 31 genes whose open reading frames were divided due to frameshifts in this, but not in the other, strains; a number of these latter genes were acid-regulated in the 26695 or J99 isolates. Another aim is to obtain a whole genome view of *H. pylori* evolution as individual human hosts progress from a normal gastric mucosa to precancer lesions (i.e. chronic atrophic gastritis). We will use the 454 FLX pyrosequencing platform to obtain deep draft whole genome sequences of *H. pylori* isolates obtained from serial endoscopies of individuals enrolled in a population-based endoscopy study [32]. *H. pylori*-infected individuals who progressed from having a normal gastric mucosa to atrophic gastritis during the 4-year interval between endoscopic examinations will be compared to sex -and age-matched control patients who did not progress (whose mucosa remained histologically normal). The relatively small size of the *H. pylori* genome will make 454 FLX pyrosequencing of multiple *H. pylori* isolates technically and economically feasible.

In addition to bacterial sequencing and comparative genomics, the next-generation sequencing platforms will provide us with a tool to study transcriptional expression and small RNA profiles of *H. pylori*. Such studies have been performed in *Salmonella*, where deep sequencing analysis of small noncoding RNA revealed a number of novel small RNAs associated with a post-transcriptional regulator, Hfq, that controls almost 20% of all salmonella genes [33].

A list of applications that will raise new challenges for experimental design and interpretation of results has recently been described [1,2,5]. However, the large amount of data generated by these instruments should lead to biologically meaningful insights and, hopefully, to clinical strategies and novel treatment options for *H. pylori*-infected individuals.

Acknowledgements

W. Paulander and P. Lehours are thanked for their helpful comments regarding the manuscript.

Transparency Declaration

The authors declare no conflicts of interest.

References

1. Shendure J, Ji H. Next-generation DNA sequencing. *Nat Biotechnol* 2008; 26: 1135–1145.
2. Mardis ER. Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet* 2008; 9: 387–402.
3. Margulies M, Egholm M, Altman WE *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 2005; 437: 376–380.
4. Ronaghi M, Karamohamed S, Pettersson B, Uhlen M, Nyren P. Real-time DNA sequencing using detection of pyrophosphate release. *Anal Biochem* 1996; 242: 84–89.
5. Rothberg JM, Leamon JH. The development and impact of 454 sequencing. *Nat Biotechnol* 2008; 26: 1117–1124.
6. Andries K, Verhasselt P, Guillemont J *et al.* A diarylquinoline drug active on the ATP synthase of *Mycobacterium tuberculosis*. *Science* 2005; 307: 223–227.
7. Oh JD, Kling-Backhed H, Giannakis M *et al.* The complete genome sequence of a chronic atrophic gastritis *Helicobacter pylori* strain: evolution during disease progression. *Proc Natl Acad Sci USA* 2006; 103: 9999–10004.
8. Andersson AF, Lindberg M, Jakobsson H, Backhed F, Nyren P, Engstrand L. Comparative analysis of human gut microbiota by barcoded pyrosequencing. *PLoS One* 2008; 3: e2836.
9. Turcatti G, Romieu A, Fedurco M, Tairi AP. A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis. *Nucleic Acids Res* 2008; 36: e25.
10. Adessi C, Matton G, Ayala G *et al.* Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms. *Nucleic Acids Res* 2000; 28: E87.
11. Dressman D, Yan H, Traverso G, Kinzler KW, Vogelstein B. Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proc Natl Acad Sci USA* 2003; 100: 8817–8822.
12. Shendure J, Porreca GJ, Reppas NB *et al.* Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* 2005; 309: 1728–1732.
13. Josenhans C, Beier D, Linz B, Meyer TF, Suerbaum S. Pathogenomics of *Helicobacter*. *Int J Med Microbiol* 2007; 297: 589–600.
14. Blaser MJ. The biology of cag in the *Helicobacter pylori*–human interaction. *Gastroenterology* 2005; 128: 1512–1515.
15. Lu H, Yamaoka Y, Graham DY. *Helicobacter pylori* virulence factors: facts and fantasies. *Curr Opin Gastroenterol* 2005; 21: 653–659.
16. Alm RA, Ling LS, Moir DT *et al.* Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. *Nature* 1999; 397: 176–180.
17. Tomb JF, White O, Kerlavage AR *et al.* The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature* 1997; 388: 539–547.
18. Falush D, Wirth T, Linz B *et al.* Traces of human migrations in *Helicobacter pylori* populations. *Science* 2003; 299: 1582–1585.
19. Salama N, Guillemin K, McDaniel TK, Sherlock G, Tompkins L, Falkow S. A whole-genome microarray reveals genetic diversity among *Helicobacter pylori* strains. *Proc Natl Acad Sci USA* 2000; 97: 14668–14673.
20. Gressmann H, Linz B, Ghai R *et al.* Gain and loss of multiple genes during the evolution of *Helicobacter pylori*. *PLoS Genet* 2005; 1: e43.
21. Moodley Y, Linz B, Yamaoka Y *et al.* The peopling of the Pacific from a bacterial perspective. *Science* 2009; 323: 527–530.
22. Hogg JS, Hu FZ, Janto B *et al.* Characterization and modeling of the *Haemophilus influenzae* core and supragenomes based on the complete genomic sequences of Rd and 12 clinical nontypeable strains. *Genome Biol* 2007; 8: R103.
23. Ehrlich GD, Hu FZ, Shen K, Stoodley P, Post JC. Bacterial plurality as a general mechanism driving persistence in chronic infections. *Clin Orthop Relat Res* 2005; Aug(437): 20–24.
24. Hiller NL, Janto B, Hogg JS *et al.* Comparative genomic analyses of seventeen *Streptococcus pneumoniae* strains: insights into the pneumococcal supragenome. *J Bacteriol* 2007; 189: 8186–8195.
25. Huse SM, Dethlefsen L, Huber JA, Welch DM, Relman DA, Sogin ML. Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing. *PLoS Genet* 2008; 4: e1000255.
26. Dethlefsen L, Huse S, Sogin ML, Relman DA. The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16S rRNA sequencing. *PLoS Biol* 2008; 6: e280.
27. Keijser BJ, Zaura E, Huse SM *et al.* Pyrosequencing analysis of the oral microflora of healthy adults. *J Dent Res* 2008; 87: 1016–1020.
28. Baltrus DA, Guillemin K, Phillips PC. Natural transformation increases the rate of adaptation in the human pathogen *Helicobacter pylori*. *Evolution* 2008; 62: 39–49.
29. Enroth H, Kraaz W, Engstrand L, Nyren O, Rohan T. *Helicobacter pylori* strain types and risk of gastric cancer: a case–control study. *Cancer Epidemiol Biomarkers Prev* 2000; 9: 981–985.
30. Bjorkman J, Andersson DI. The cost of antibiotic resistance from a bacterial perspective. *Drug Resist Updat* 2000; 3: 237–245.
31. Bjorkholm B, Sjolund M, Falk PG, Berg OG, Engstrand L, Andersson DI. Mutation frequency and biological cost of antibiotic resistance in *Helicobacter pylori*. *Proc Natl Acad Sci USA* 2001; 98: 14607–14612.
32. Aro P, Storskrubb T, Ronkainen J *et al.* Peptic ulcer disease in a general adult population: the Kalixanda study: a random population-based study. *Am J Epidemiol* 2006; 163: 1025–1034.
33. Sittka A, Lucchini S, Papenfort K *et al.* Deep sequencing analysis of small noncoding RNA and mRNA targets of the global post-transcriptional regulator, Hfq. *PLoS Genet* 2008; 4: e1000163.